Final Project Possibilities

Adam Meyers
New York University
Summary

• Possible Extensions of HW7
• TAC
• Downloadable Tasks/ Corpus Data
• Existing NLP tools and Packages
• Annotation Projects
• Research Papers
Getting Existing Resources

• I will outline several possibilities, some of which depend on existing resources
• Some resources are readily available
• Other resources require me to do something
  – If something interests you, I will make it a priority
  – I will make some materials available regardless
• You can also use the resources made available for the class so far.
• You can also find other resources or create your own
How We Derived Homework 7

• Several Nombank noun classes lend themselves to a relation task.
  – % belongs to one such class
  – % happens to have many instances

• Hypothesized Properties of Good Relation Classes:
  – Many instances
  – Only 1 or 2 “core” arguments
  – Predicates should be semantically similar to each other
  – Adjuncts Modifiers (ARGMs) should be rare and not too diverse
Same Class as % and Similar Classes

• Partitive-Quant (the same class as %)
  – Includes over 200 distinct words: %, percent, 1/10, acre, array, jumble, quart, wave, watt, yen, …
    • Newsstands are packed with a colorful array of magazines/ARG1.

• Partitive-Part – not transparent, necessarily a part of ARG1
  – Over 200 words including: anther, back, borough, component, …
    • Of the 11 components to the index/ARG1, only three others rose in September
  – Also includes meronyms – idiosyncratic part-of cases
    • … stronger roofs for light trucks/ARG1 and minivans/ARG1

• Group – like Partitive-Quant, except can have ARG2 (leader)
  – Over 60 words including: army, association, band, clan, troop, …
    • Robin/ARG2's band of merry men/ARG1

• Share – like Partitive-Part, except ARG0 (owner) and ARG2 (value)
  – 13 words including: allotment, chunk, interest, niche, portion, …
    • Time Warner/ARG0's 50%/ARG2 interest in Working Women/ARG1 and Working Mother/ARG1
Complications: Multiple Instances of the Same Argument

• Conjunction
  – We will assume that conjoined phrases have multiple heads and both should be marked if the phrase is an argument
    • *Time Warner/ARG0 's 50 %/ARG2 interest in Working Women/ARG1 and Working Mother/ARG1*
    • … stronger *roofs* for light *trucks/ARG1 and minivans/ARG1*

• Secondary Themes
  – ARG1 for these classes can sometimes be duplicated (as ARG3)
    • *Tiny neck/ARG1 arteries of rats/ARG3*
    • *the plane/ARG1 's wing/ARG3 flaps*
  – For scoring purposes ARG1 and ARG3 are equivalent
More About ARG1 for Partitive/Group/Share Nouns

- ARG1 can be an NP or PP that lists set members or whole object that head noun is a part of
  - Robin/ARG2's band of merry men/ARG1
  - the 11 components to the index/ARG1
  - The price/ARG1 rose/SUPPORT 5 %

- ARG1 (or ARG3) can describe a property of a set
  - team of managers/ARG1; management/ARG1 team
  - team of editors/ARG1, editing/ARG1 team
  - 5-editor/ARG1 management/ARG3 team
Arguments for PARTITIVE/GROUP/SHARE nouns

• ARG1:
  – Partitive-Quant/Group/Share (Quantified)
  – Partitive-Part (Whole)

• ARG0:
  – Share (Owner)

• ARG2:
  – Share (Value)
  – Group (Employer/Leader/Pivot)

• ARG3 (Secondary theme) ≈ ARG1
Tasks Based on These Data

• Data Like the % task except:
  – Possible Multiple ARG1s
  – Possible other ARGs
  – More Data
  – More Diverse predicates

• Modified HW7 Systems Could Do this Task
  – You may want to simplify the task
    • i.e., remove problematic cases
  – System could deal with ARG0, ARG2, ARG3, multiple ARGs

• If task is simplified encode simplification in scorer
  – We will have a test set in the same format as Dev data
  – If your system and scorer works on the dev data, it should work on the test data
TAC2010 Knowledge-based Population Task

- Website: http://nlp.cs.qc.cuny.edu/kbp/2010/
  - Task definition documents:
    - http://projects.ldc.upenn.edu/kbp/specifications/TAC_2010_KBP_Slot_Filling_Annotation_Guidelines_V1.2.pdf

- Extracting Info about People and Organizations
  - Like Wikipedia Info Boxes
  - Pick a small subset of the slots to fill based on the specs (or your own modified specs)
  - Build your own small corpus, just marking those relations.
    - Use most of the corpus as training, but keep at least 100 examples as your development and 100 for testing.
Write a system that fills a few slots from text

- 26 Possible Slots for People: Alternative Names, Time of Birth, Time of Death, Name of Spouse, Names of Siblings, Age, Country of birth, ...
- 16 Possible Slots for Organization: Alternative Names, Religious/Political Affiliation, Top members/employees, founded by, ...

A Possible Method

- Find People/Organization in text
  - Use NE tagger (I will make one available)
- Find keywords indicating slot fill possibility
  - Manually create set of keywords
- Fill slot fill: regular expression, Maxent classifier, or combination
The Wall Street Journal Penn Treebank

• I could make this available in several forms
• Upenn: Trees, POS tags, Noun Groups
• BBN:
  – NE – Inline annotation – you can convert it to one token per line or use it as is
  – Coreference – marks pronoun coreference – there is probably a bit of corpus preparation to make this work
Downloadable Tasks with Annotated Corpora (for Testing and/or Training)

• Corpora for Drug-Drug Interaction
  – http://labda.inf.uc3m.es/DrugDDI/DrugDDI.html

• WePS – searching for entities on the Web

• Spanish Corpus with POS tags
  – http://www.comp.lancs.ac.uk/linguistics/crater/spanishfiles.html
Some Downloadable Corpora

- The Open American National Corpus
  - [http://www.anc.org/OANC/](http://www.anc.org/OANC/)
  - A variety of different types of data
  - A limited amount of manually annotated data
  - Automatically annotated data from various programs
  - Most annotation is offset annotation

- The Singapore SMS corpus

- Wikipedia
  - Controversial Articles in Japanese and English
  - Wikipedia XML
Tools and Packages

• You can use output of NLP tools, as input to your system, provided you make your contribution clear.

• Examples:
  – The Maxent Package
  – Tools within NLTK
  – I will make a version of JET's NE tagger available
  – Other tools you download
Sample Annotation Projects

- Apply one a known type of Annotation (NE, POS, Chunking, semantic role labeling) to a new domain of text: web data, technical data, etc.
- Develop specs and annotate new classes of NEs, Relations, or Events
- Develop specs and annotate an interesting phenomenon:
  - quantifier scope
  - sentiment (your version)
  - Idiomatic expressions
Survey Papers

• If you select an area, We can recommend some articles/books to get you started
• Research Areas: Machine Translation, Summarization, Question Answering, Sentiment Analysis, Information Extraction, Reference Resolution, Predicate Argument Structure, …